Integrating statistical prior knowledge into convolutional neural networks

Fausto Milletari, Alex Rothberg, Jimmy Jia, Michal Sofka

4Catalyzer Corporation

Abstract. In this work we show how to integrate prior statistical knowledge, obtained through principal components analysis (PCA), into a convolutional neural network in order to obtain robust predictions even when dealing with corrupted or noisy data. Our network architecture is trained end-to-end and includes a specifically designed layer which incorporates the dataset modes of variation discovered via PCA and produces predictions by linearly combining them. We also propose a mechanism to focus the attention of the CNN on specific regions of interest of the image in order to obtain refined predictions. We show that our method is effective in challenging segmentation and landmark localization tasks.

1 Introduction and Related Work

In the past decade multiple authors proposed approaches to perform tasks such as medical image segmentation [1,4,12,14] and registration [3] using PCA.

When representing shapes through a fixed number of control points, PCA can be used to build a point distribution model (PDM) by finding the principal modes of variation of the shapes across the training dataset. A segmentation algorithm can then rely on both image data and prior knowledge to fit a contour that is in agreement with the shape model. The resulting segmentation is anatomically correct, even when the image data is insufficient or unreliable because of noise or artifacts. These approaches are referred to as active shape models (ASM) in literature [5] and were shown to be applicable to a variety of problems. For example in [1], a hardly visible portion of the brain, imaged by ultrasound through the temporal bone window of the skull, was reliably segmented using a 3D active contour.

Several other approaches unite the advantages brought by active shape models with active appearance models. In [12], volumetric ultrasound and MRI images of the heart were segmented using 3D active appearance models. A common shortcoming of these approaches is the difficulty to define an energy function to optimize such that a contour evolves correctly and appropriately segments the region of interest after a few hundred iterations of an optimization algorithm.

More recent approaches, mainly based on machine learning, have taken advantage of implicit prior knowledge and advanced handcrafted or learned features in order to overcome the limitations of previous, optimization-based techniques. In [11], a random Hough forest was trained to localize and segment the left

ventricle of the heart. The notion of shape model was enforced through the constraints imposed by the voting and segmentation strategy which relied on re-projecting portions of the ground truth contours encountered during training onto previously unseen examples. This idea was later extended in [8].

Deep learning-based approaches have been recently applied to medical image analysis. Segmentation architecture leveraging a fully convolutional neural network was proposed to process 2D images [13] and volumes [2,10]. These methods do not make use of any statistical shape model and rely only on the fact that the large receptive field of the convolutional neural network will perceive the anatomy of interest all at once and therefore improbable shapes will be predicted only rarely in modalities such as MRI and microscopy images. An interesting approach [7,9] fusing Hough voting with CNNs was applied to ultrasound images and MRI brain scans. Although the Hough-CNN delivered accurate results, its design prevents end-to-end training.

In this work we propose to include statistical prior knowledge obtained through PCA into a deep neural convolutional network. Our PCA layer incorporates the modes of variation of the data at hand and produces predictions as a linear combination of the modes. This process is used in a procedure that focuses the attention of the subsequent CNN layers on the specific region of interest to obtain refined predictions. Importantly, the network is trained end-to-end with the shape encoded in a PCA layer and the loss imposed on the final location of the points. In this way, we want to overcome the limitations of previous deep learning approaches which lack strong shape priors and the limitations of active shape models which miss advanced pattern recognition capabilities. Our approach is fully automatic and therefore differs from most previous methods based on ASM which require human interaction. The network outputs the prediction in a single step without requiring any optimization loop.

We apply our method to two challenging ultrasound image analysis tasks. In the first task, the shape modeling improves the accuracy of the landmark localization in 2D echocardiography images acquired from the parasternal long axis view (PLA). In the second task, the algorithm improves the dice coefficient of the left ventricle segmentation masks on scans acquired from the apical two chamber view of the heart.

2 Method

We are given a training set containing N images $I = \{I_1, \ldots, I_N\}$ and the associated ground truth annotations $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, $\mathbf{y}_i \in \mathbb{R}^{2P}$ consisting of coordinates referring to P key-points which describe the position of landmarks. We use the training set to first obtain the principal modes of variation of the coordinates in Y and then train a CNN that leverages it. In order to contrast the loss of fine-grained details across the CNN layers, we propose a mechanism that focuses the attention of the network on full-resolution details by cropping portions of the image in order to refine the predictions (Figure 1 and 2). Our

architecture is trained end-to-end, and all the parameters of the network are updated at every iteration.

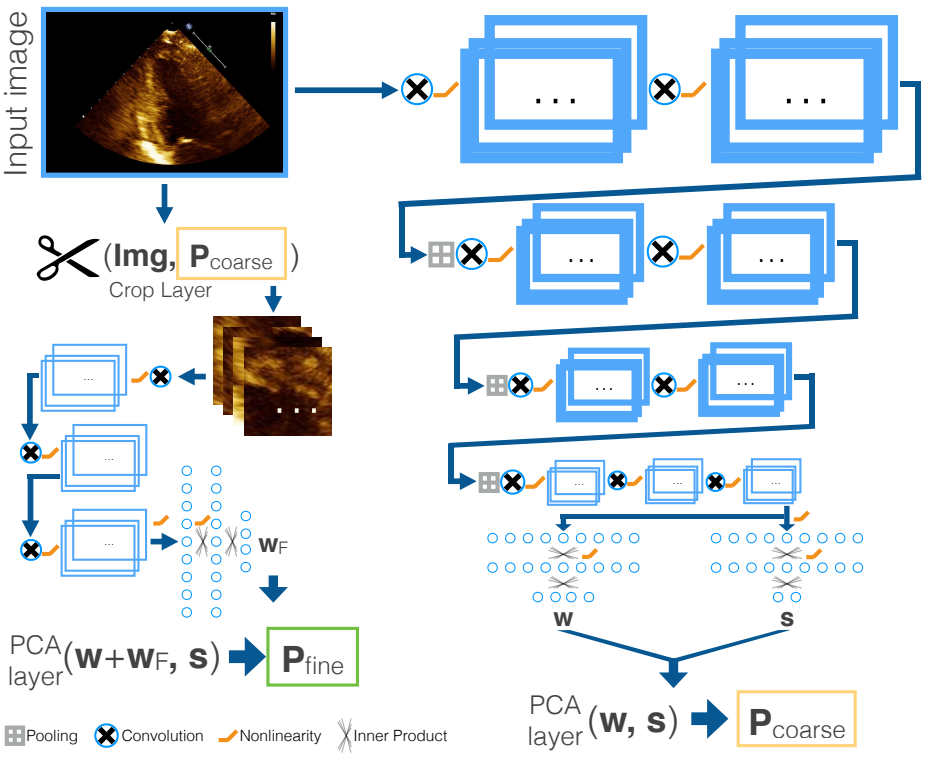

Fig. 1. Schematic representation of the proposed network architecture.

2.1 Building a shape model through PCA

Much of the variability of naturally occurring structures, such as organs and anatomical details of the body, is not arbitrary: symmetries and correlations exist between different shape portions or anatomical landmarks. Principal component analysis (PCA) [15] can be used to discover the principal modes of variation of the dataset at hand. When we describe shapes as aligned points sets across the entire dataset, PCA reveals what correlations exist between different points and defines a new coordinates frame where the principal modes of variation correspond to the axes. First, we subtract mean of each shape point in every shape $\mathbf{y}_i$ as

$$\tilde{\mathbf{y}}_\mathbf{i} = \mathbf{y}_i - \mu, \text{ with } \mu = \frac{1}{N}\sum_i \mathbf{y}_i. \tag{1}$$

We then construct matrix $\tilde{\mathbf{Y}}$ all samples in our dataset by stacking $\{\mathbf{y}_i\}$ column-wise. Finally, we compute the eigenvectors of the covariance matrix $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top$. This corresponds to $\mathbf{U}$ in

$$\tilde{\mathbf{Y}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \tag{2}$$

which is obtained via singular value decomposition (SVD). The matrix $\boldsymbol{\Sigma}$ is diagonal and contains elements $\{\sigma_1^2, \ldots, \sigma_K^2\}$ which are the eigenvalues of the covariance matrix and represent the variance associated with each principal component in the eigenbase.

Any example in the dataset can be synthesized as a linear combination of the principal components.

$$\mathbf{y}_i = \mathbf{U}\mathbf{w} + \mu \tag{3}$$

Each coefficient of the linear combination governs not only the position of one, but multiple correlated points that, in our case, describe the shape at hand. Imposing constraints on the coefficients weighting the effect of each principal component, or reducing their number until the correct balance between percentage of retained variance and number of principal components is reached, it is possible to synthesize shapes that respect the concept of "legal shape" introduced before.

2.2 Network architecture

In this work we use a CNN, schematically represented in Figure 1, to perform predictions using the principal components stored in the matrix $\mathbf{U}$.

We do not train the CNN to perform regression on the weights $\mathbf{w}$ in Equation 3, but we resort to an end-to-end architecture instead: the network directly uses the PCA eigenbase to make predictions $\tilde{\mathbf{y}}_\mathbf{i} \in \mathbb{R}^{2P}$ from an image $\mathbf{I}_\mathbf{i}$ in form of key-points locations. This has direct consequences on the training process. The network learns, by minimizing the loss $l = \sum_i \|\tilde{\mathbf{y}}_i - \mathbf{y}_i\|_2^2$, to steer the coefficients while being "aware" of their effect on the results. Each of the weighs w_j controls in fact the location of multiple correlated key-points simultaneously. Since the predictions are obtained as a liner combination of the principal components, they obey the the concept of "legal shape" and therefore are more robust to missing data, noise and artifacts.

Our network comprises two branches. The first employs convolutional, pooling and fully connected layers, and produces a coarse estimate of the key-point locations via PCA. The second operates on full resolution patches cropped from the input image around the coarse key-point locations. The output of the second network refines the predictions made by the first by using more fine-grained visual information. Both the branches are trained simultaneously and are fully differentiable. The convolutions are all applied without padding and they use kernels of size 3×3 in the first CNN branch and 5×5 in the second, shallower, branch. The nonlinearities used throughout the network are rectified linear functions. All the inputs of the PCA layer, are not processed through nonlinearities.

Our PCA layer implements a slightly modified version of the synthesis equation in 3. In addition to the weights $\mathbf{w}$, which are supplied by a fully connected layer of the network, we also provide a global shift $\mathbf{s}$ that is applied to all the predicted points. Through the bi-dimensional vector $\mathbf{s}$ we are able to cope with translations of the anatomy of interest. With a slight abuse of notation we can therefore re-write the modified Equation 3 as

$$\mathbf{y}_i = \mathbf{U}\mathbf{w} + \mu + \mathbf{s}. \tag{4}$$

The layer performing cropping follows an implementation inspired to spatial transformers [6] which ensures differentiability. A regular sampling pattern is translated to the coarse key-point locations and the intensity values of the surrounding area are sampled using bilinear interpolation. Having P key-points we obtain P patches for each of the K images in the mini-batch. The resulting KP patches are then processed through a 3-layers deep convolutional neural network using 8 filters applied without padding, which reduces their size by a total of 12 pixels. After the convolutional layers the patches are again arranged into a batch of K elements having $P \times 8$ channels, and further processed through three fully connected layers, which ultimately compute $\mathbf{w}_A$ having the same dimensionality of $\mathbf{w}$. The refined weights $\mathbf{w}_F$ which are employed in the PCA layer to obtain a more accurate key-point prediction, are obtained as $\mathbf{w}_F = \mathbf{w}_A + \mathbf{w}$.

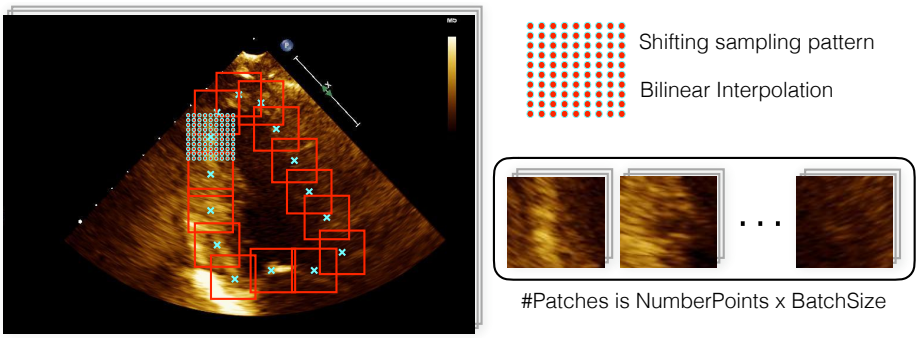

Fig. 2. Schematic representation of the crop layer. The shifting sampling pattern is centred at the landmark positions. High resolution patches are cropped from the input image and organized in a batch.

3 Results

We tested our approach on two different ultrasound dataset depicting the human heart. Our aim was to solve two different tasks. The first task is segmentation of the left ventricle (LV) of the heart form scans acquired from the apical view,

while the second task is a landmark localization problem where we aim to localize 14 points of interest in images acquired from the parasternal long axis view. In the first case our model leverages prior statistical knowledge relative to the shape of the structures of interest, while in the second case our model captures the spatiotemporal relationships between landmarks across cardiac cycles of different patients. For the segmentation task we employ a total of 1100 annotated images, 953 for training and 147 for testing. The landmark localization task was performed on a test set of 47 images by a network trained on 706 examples. The total number of annotated images employed for the second task was therefore 753. There was no overlap between the training and test patients. All the annotations were performed by expert clinicians specifically hired for this task.

Our python implementation relies on the popular Tensorflow framework. All experiments have been performed on standard PC equipped with a Nvidia Tesla K80 GPU, with 12 GB of video memory, 16 GB of RAM and a 4 Cores Intel Xeon CPU running at 2.30 GHz. Processing a single frame took a fraction of a second.

3.1 Segmentation

We represent the shapes of interest as a set of 32 corresponding key-points which are interpolated using a periodic third degree B-spline. The result is a closed curve delineating the left ventricle of the heart. We compare our results with:

- CNN with a structure similar to the one of the main branch of our architecture, which does not employ a PCA layer but simply regresses the positions of the landmarks without imposing further constraints.
- The U-Net architecture [13], which predicts segmentation masks having values comprised in the interval 0, 1which are then thresholded at 0.5.

We train all the architectures for 100 epochs, ensuring in this way convergence The results are summarized in Table 1.

Table 1. Summary of the results obtained for the segmentation task.

	Dice Score		
Architecture	Mean	Min	Max
Proposed	$0.87 \pm \mathbf{0.041}$	**0.80**	**0.96**
CNN	0.86 ± 0.042	0.78	0.93
U-Net	0.88 ± 0.063	0.63	**0.96**

In Figure 3 we report the distribution of Dice scores obtained on the test set in form of histogram.

3.2 Landmark localization

The results of the landmark localization task are presented in Table 2. The shape modeling PCALayer introduces constraints that help improve accuracy of the

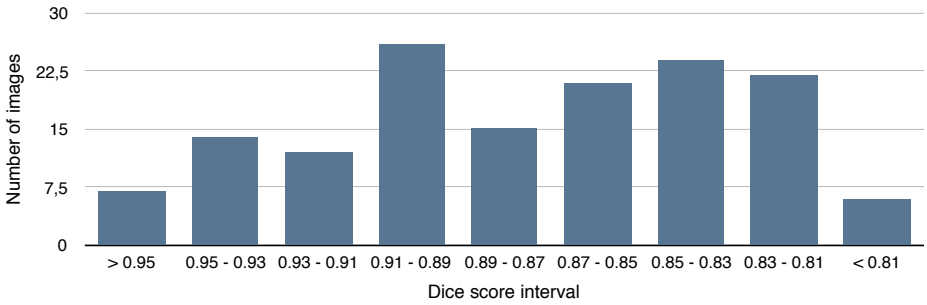

Fig. 3. Distribution of Dice Scores on the test set.

measurements. Compared to the convolutional architecture with fully connected layers regressing the point locations, the explicit shape constraints better guide the relative displacement of the individual measurement points.

Table 2. Summary of the results obtained for the landmark localization task.

	Distances in mm		
Architecture	Mean	Min	Max
Proposed	2.06 ± 1.89	0.01	10.46
CNN	2.33 ± 1.67	0.15	8.78

4 Conclusion

We proposed a method to incorporate prior shape constraints into deep neural networks. This is accomplished by a new Principal Component Analysis (PCA) layer which computes predictions from linear combinations of modes of shapes variation. The predictions are used to steer the attention of the subsequent convolutional layers to refine the prediction estimates.

The proposed architecture improves the robustness and accuracy of the segmentation results and multiple measurements. Our experiments on the left ventricle ultrasound scans in a two-chamber apical view showed higher minimum dice coefficients (fewer failures and lower standard deviation) than a CNN architecture regressing the point locations and a U-Net architecture predicting the foreground probability map. Our results on multiple measurements of heart structures in the parasternal long axis view show lower measurement errors.

References

1. Ahmadi, S.A., Baust, M., Karamalis, A., Plate, A., Boetzel, K., Klein, T., Navab, N.: Midbrain segmentation in transcranial 3d ultrasound for parkinson diagnosis.

In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 362–369. Springer (2011)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer (2016)
3. Cootes, T.F., Beeston, C., Edwards, G.J., Taylor, C.J.: A unified framework for atlas matching using active appearance models. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 322–333. Springer (1999)
4. Cootes, T.F., Edwards, G.J., Taylor, C.J., et al.: Active appearance models. IEEE Transactions on pattern analysis and machine intelligence 23(6), 681–685 (2001)
5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding 61(1), 38–59 (1995)
6. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
7. Kroll, C., Milletari, F., Navab, N., Ahmadi, S.A.: Coupling convolutional neural networks and hough voting for robust segmentation of ultrasound volumes. In: German Conference on Pattern Recognition. pp. 439–450. Springer (2016)
8. Milletari, F., Ahmadi, S.A., Kroll, C., Hennersperger, C., Tombari, F., Shah, A., Plate, A., Boetzel, K., Navab, N.: Robust segmentation of various anatomies in 3d ultrasound using hough forests and learned data representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 111–118. Springer (2015)
9. Milletari, F., Ahmadi, S.A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Boetzel, K., et al.: Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound. arXiv preprint arXiv:1601.07014 (2016)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv preprint arXiv:1606.04797 (2016)
11. Milletari, F., Yigitsoy, M., Navab, N.: Left ventricle segmentation in cardiac ultrasound using hough-forests with implicit shape and appearance priors
12. Mitchell, S.C., Bosch, J.G., Lelieveldt, B.P., Van der Geest, R.J., Reiber, J.H., Sonka, M.: 3-d active appearance models: segmentation of cardiac mr and ultrasound images. IEEE transactions on medical imaging 21(9), 1167–1178 (2002)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
14. Sofka, M., Wetzl, J., Birkbeck, N., Zhang, J., Kohlberger, T., Kaftan, J., Declerck, J., Zhou, S.: Multi-stage learning for robust lung segmentation in challenging CT volumes. In: Proceedings of the 14th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2011). Toronto, Canada (18–22 Sep 2011)
15. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems 2(1-3), 37–52 (1987)